MULTIDIMENSIONAL DETECTION OF OUTLIERS IN CLINICAL REGISTERS

Branislav Šiška

Master Degree Programme (2), FEEC BUT

E-mail: xsiska09@stud.feec.vutbr.cz

Supervised by: Daniel Schwarz

E-mail: schwarzd@feec.vutbr.cz

Abstract: Incorrect data in clinical registers can lead to inaccurate or wrong results. This project is aimed at monitoring and evaluation of data in clinical registers. Usual methods to identify incorrect data are one-dimensional statistical methods per each variable in the register. Proposed method finds outliers in data using machine learning combined with multidimensional statistical methods that transform all column variables of clinical register to one, representing one record of a patient in the register.

Keywords: clinical register, detection of outliers, data fraud, machine learning

1 ÚVOD

Klinický register je systém jasne definovaných zdravotných, alebo demografických údajov od pacientov so špecifickými zdravotnými charakteristikami, ktoré sú uchované v centrálnej databáze. Klinické registre môžu slúžiť ako monitorovací nástroj na zlepšenie lekárskej starostlivosti, alebo k poskytovaniu informácií o ľuďoch s daným ochorením. Podľa usporiadania konkrétneho registra z neho môžu čerpať informácie pacienti, lekári aj vedci. Dáta v klinickom registri sa zbierajú prostredníctvom formulárov. Pri vyplňovaní formulárov sú často údaje o pacientoch zadané nepresne, alebo sú niektoré polia vynechané. To spôsobuje nepresnosti pri výpočtoch, čo môže viesť až ku nesprávnym záverečným výsledkom klinickej štúdie [1]. Systematické anomálie vznikajú dôsledkom chýb zberného programu, nejasne definovaných parametrov, alebo zneužitia zberu dát. Náhodné chyby vznikajú nepresným prepisom dát, alebo preklepmi pri ich zadávaní. Navrhnutá metóda slúži na zistenie odľahlých hodnôt pri zadávaní dát do databázy klinických registrov. Odhalí chyby zberného programu, chyby vzniknuté z nepozornosti a vďaka multidimenzionálnemu prístupu aj cielené zadávanie fiktívnych hodnôt. Tieto hodnoty sú veľmi ťažko odhaliteľné, pretože sú vo fyziologickom rozmedzí.

2 PREDSPRACOVANIE DÁT

2.1 EXTRAKCIA PREMENNÝCH

Reálne dáta z klinických registrov je nutné pred ďalšou analýzou upraviť. Obsahujú totiž premenné, ktoré nie sú pre následnú analýzu potrebné. Keďže analyzované vstupné dáta dosahovali dostatočnú veľkosť, pre nevyplnenú jednu a viac hodnôt z premenných boli odstránené odpovedajúce pacientske záznamy. Táto metóda je najjednoduchšia, ale treba si uvedomiť, že sa jej použitím stráca určitá informácia a preto treba byť pri jej použití opatrný. Pri analyzovaných súboroch dát menšieho rozsahu sa odporúča chýbajúce hodnoty doplniť priemerom hodnôt, ktoré sú pre danú premennú k dispozícii. Ďalšia z dostupných doplňovacích metód je metóda mnohonásobného regresného modelu.

2.2 TRANSFORMÁCIA NA NUMERICKÝ VEKTOR

Extrakcia potrebných premenných z pôvodného formulára je pre nenumerické premenné nasledovaná ich transformáciou na numerický vektor. Premenné obsahujúce dátumové hodnoty sú transformované na numerický vektor ako počet dní uplynutých od 0. januára roku 0. Pre kvantifikovateľne vymenované premenné sa z usporiadaných hodnôt vytvorí očíslovaná škála pre každú hodnotu danej premennej. Nekvantifikovateľné premenné sa prevedú na binárne premenné.

2.3 ŠTANDARDIZÁCIA

Keďže každá zo skúmaných premenných môže pochádzať z iného rozloženia a rozsahu, ďalším krokom procesu je štandardizácia premenných ich rozpätím na hodnoty od nula do jedna. Na konci procesu predspracovania dát tak každý riadok vybraného formulára klinického registru predstavuje jeden plne vyplnený pacientsky záznam a každý odpovedajúci stĺpec formulára predstavuje numericky reprezentovanú záujmovú premennú .

3 REALIZÁCIA METÓDY

3.1 MODIFIKOVANÝ K-MEANS

Pre každú premennú z formulára registra je vytvorený jeden centroid. Od tohto centroidu sa následne spočíta podobnosť pre každý pacientsky záznam daného formulára. Jeden riadok záznamu v registri, tak už nie je reprezentovaný súborom premenných, ale jednou hodnotou vybranej podobnostnej metriky. Na klasifikáciu podobnosti sa použili tri metriky. Euklidova vzdialenosť ako základná podobnostná metrika založená na Pytagorovej vete. Z dôvodu štandardizácie dát bola zvolená aj kosínová vzdialenosť, ktorá je definovaná ako kosínová podobnosť odpočítaná od 1. Ako tretia bola zvolená Mahalanobisova vzdialenosť, ktorá berie v úvahu koreláciu medzi jednotlivými parametrami a rešpektuje rozdielnu variabilitu dát. Všetky vyššie spomenuté podobnostné metriky sú rozobrané podrobnejšie v [2].

3.2 URČENIE PRAHU

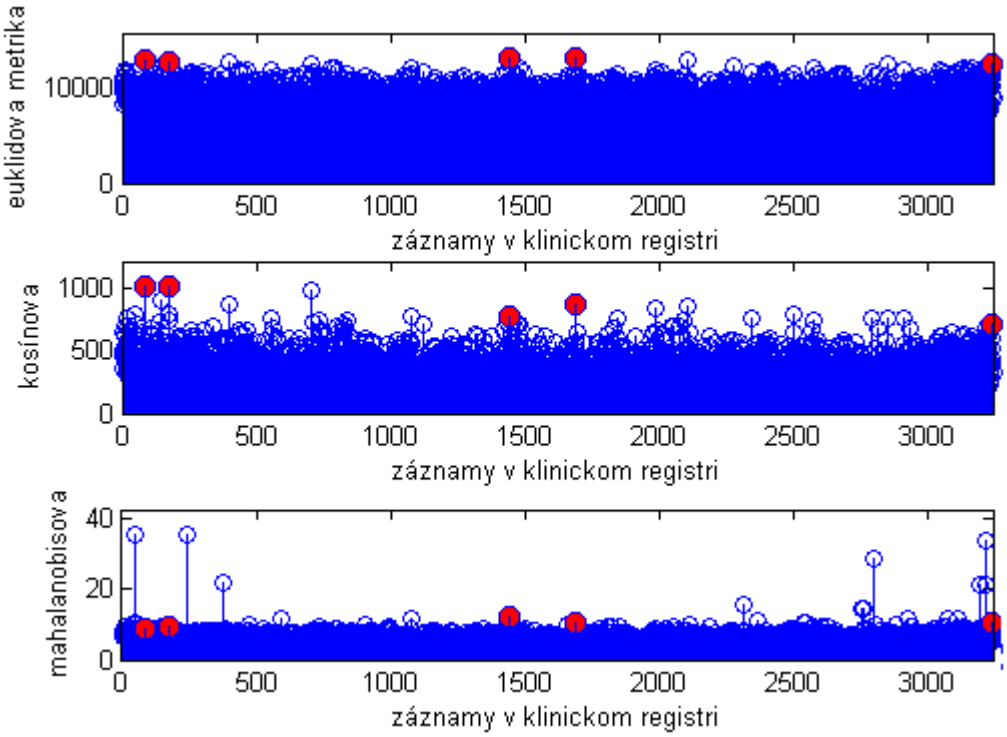
Na určenie hranice podobnosti, za ktorou sa daný záznam v klinickom registri bude označovať za odľahlý sa testovalo viacero metód. Najlepšie výsledky dosahovala hodnota hranice určená ako 1,5 násobok interkvartilového rozpätia pripočítaná k tretiemu kvartilu analyzovanej premennej [3]. Ďalšia z možných metód určenia prahu je založená na detekcii zlomu v krivke závislostí počtu odľahlých hodnôt na hodnote prahu. Tento prah je iterovaný automaticky od minima do maxima danej podobnostnej metriky s konfigurovateľným krokom iterácie. Posledná z testovaných možností na zistenie optimálneho prahu určí za odľahlé hodnoty tie záznamy, ktoré sa v histograme vzdialeností vyskytovali do desiateho percentilu, alebo nad deväťdesiatym percentilom skúmanej vzdialenosti.

3.3 POTVRDENIE DETEKCIE ODĽAHLEJ HODNOTY

Na prehlásenie záznamu v klinickom registri za vysoko odľahlý ho musia za odľahlý zaradiť všetky tri podobnostné metriky. Ak záznam zaradili za odľahlý len dve metriky, ide o stredne odľahlý záznam. Vysoko a stredne odľahlé záznamy by mali byť následne z klinického registru odstránené. Pri zaradení záznamu za nízko odľahlý jednou metrikou je nutné potvrdenie, či sa skutočne jedná o odľahlú hodnotu potrebné schváliť ručne. Pri potvrdení je možné si prehliadnuť doplňujúce informácie o danom zázname v odpovedajúcom formulári. Zobrazuje sa jednoznačný identifikátor pacientskeho záznamu, hodnoty a zobrazenie samotných podobnostných kritérií. Porovnanie troch podobnostných kritérií pre skúmaný formulár je na obrázku 1. Osa y predstavuje 3241 pacientskych záznamov z formulára klinického registra a osa x predstavuje vzdialenosť centroidu od odpovedajúceho pacientskeho záznamu. Čím je hodnota tejto metriky väčšia, tým je sledovaný pacientsky záznam menej podobný centroidu. Červenou farbou sú vyznačené pacientske záznamy, ktoré boli označené za odľahlé všetkými troma podobnostnými metrikami.

3.4 SPRÁVNOSŤ METÓDY

Na testovanie správnosti zvoleného prahu sa využila ROC krivka. Pre správne testovanie je potrebné vedieť pozície skutočných odľahlých hodnôt od skúseného dátového manažéra klinického registru. Tieto hodnoty sa však nepodarilo získať. Preto boli detekované odľahlé hodnoty testované proti náhodne vygenerovaným dátam podobnostných metrík. Plocha pod ROC krivkou bola najnižšia pre Euklidovu podobnostnú metriku s hodnotou 0,65. Pre kosínovú podobnostnú metriku bola plocha pod ROC krivkou s hodnotou 0,76. Najlepšie výsledky dosiahla Mahalanobisova metrika s hodnotou 0,93.



Obrázok 1: Výsledky podobnostných metrík – červenou sú označené pacientske záznamy, ktoré boli za odľahlé určené všetkými podobnostnými metrikami - vysoko odľahlé pacientske záznamy

4 ZÁVER

Navrhnutá metóda detekuje odľahlé hodnoty záznamov v klinických registroch. S použitím multidimenzionálneho prístupu dokáže detekovať aj hodnoty fiktívneho zadávania dát vo fyziologickom rozmedzí, ktoré sú inak veľmi ťažko identifikovateľné. Tieto hodnoty skresľujú výsledky a môžu tak viesť k nesprávnym záverom štúdie. Preto je dôležité týmto hodnotám prikladať zvýšenú pozornosť. Ak sa jedná o úmyselné zneužitie dát, alebo chybu zberného programu, môžu byť zdetekované odľahlé hodnoty po kontrole oprávnenou osobou z klinického registra odstránené. Ak by sa jednalo o chybu pri zadávaní, alebo prepise dát môže sa hodnota jednoducho opraviť a odpovedajúci pacientsky záznam bude ponechaný v klinickom registri.

REFERENCIE

[1] SVOBODNÍK, Adam, Regina DEMLOVÁ a Ladislav PECEN. Klinické studie v praxi. Brno: Facta Medica, 2014. ISBN 978-80-904731-8-8.

[2] HARUŠTIAKOVÁ, Danka. Vícerozměrné statistické metody v biologii. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-791-8.

[3] Zar JH. Biostatistical Analysis. 5th edition, Pearson Prentice-Hall, New Jersey, 2010. ISBN 0321656865.